Evaluating longitudinal invariance in dimensions of mental health across adolescence: An analysis of the Social Behavior Questionnaire

Murray, Aja Louise ; Eisner, Manuel ; Obsuth, Ingrid ; Ribeaud, Denis

Abstract: Measurement invariance over time (longitudinal invariance) is a core but seldom-tested assumption of many longitudinal studies on adolescent psychosocial development. In this study, we evaluated the longitudinal invariance of a brief measure of adolescent mental health: the Social Behavior Questionnaire (SBQ). The SBQ was administered to participants of the Zurich Project on the Social Development of Children and Youths in up to four waves spanning ages 11 to 17. Using a confirmatory factor analysis approach, metric invariance held for all constructs, but there were some violations of scalar and strict invariance. Overall, intercepts tended to increase over time while residual variances decreased. This suggests that participants may become more willing or able to identify and report on certain behaviors over time. The noninvariance was not practically significant in magnitude, except for the Anxiety dimension where artifactual increases over development would be liable to occur if invariance is not appropriately modeled. Overall, results support the utility of the SBQ as an omnibus measure of psychosocial health across adolescence.

DOI: https://doi.org/10.1177/1073191117721741

Evaluating longitudinal invariance in dimensions of mental health across adolescence: An analysis of the Social Behavior Questionnaire

Aja Louise Murray, Ingrid Obsuth, Manuel Eisner, and Denis Ribeaud

Abstract

Measurement invariance over time is a core but seldom-tested assumption of many longitudinal studies on adolescent psychosocial development. In this study, we evaluated the longitudinal invariance of a brief measure of adolescent mental health: the Social Behavior Questionnaire (SBQ). The SBQ was administered to participants of the Zurich Project on the Social Development of Children and Youths (z-proso) in up to 4 waves spanning ages 11 to 17. Using a confirmatory factor analysis approach, metric invariance held for all constructs but there were some violations of scalar and strict invariance. Overall, intercepts tended to decrease over time while residual variances decreased. This suggests that participants may become more willing or able to identify and report on certain behaviours over time.

. The non-invariance was not practically significant in magnitude, except for the Anxiety dimension. If non-invariance in Anxiety is not appropriately taken into account, increases in this construct over development would be liable to occur. Overall, results support the utility of the SBQ as a measure of mental health across adolescence and suggest that across longitudinal studies in adolescence, Keywords: longitudinal invariance; Social Behavior Questionnaire; adolescence; mental health

A key goal of longitudinal research in adolescence is to illuminate processes of development through the examination of stability and change in emotional, psychological and behavioural traits. The validity of conclusions drawn from such research relies on the availability of comparable trait estimates over development. This in, turn requires at least partial longitudinal invariance, namely, that a subset of items capture the same construct on the same measurement scale over time (Edwards & Wirth, 2012; Widaman et al., 2010). Measurement invariance for items can be violated for a variety of reasons, especially in periods of substantial developmental change such as adolescence. When this happens and it is not appropriately modelled, apparent changes over time could be due to changes in the way a construct is measured; true changes could be masked; or an entirely different construct could be measured at different time points (e.g. Edwards & Wirth, 2009). However, non-invariance can also provide insights into developmental processes. It may, for example, reveal changes in the way that adolescents perceive and interpret their behaviour and symptoms in the context of their rapidly changing social and internal psychological environment. It may also reveal how the manifestations of psychosocial constructs evolve across this period of development.

Constructs of core interest in adolescent psychosocial development include conduct issues, depression, anxiety, attention deficit hyperactivity disorder (ADHD), and well as prosocial behaviour. A sizeable body of research has sought to characterise development of these dimensions in adolescence. This includes studies characterising average developmental trajectories, developmental trajectory subtypes, developmental inter-relations, and identifying predictors and outcomes of developmental trajectories (e.g. Carlo et al., 2007; Crocetti et al., 2009; Dekker et al., 2007; Luengo Kanacri et al., 2013; Martino et al., 2008; Marmorstein, 2009; Murray, Eisner & Ribeaud, 2016a; Murray, Obsuth et al., 2016a; Murray, Obsuth et al.,

2017; Nantel-Vivier et al., 2009; Van Oort et al., 2009). Studies of this kind have, for example, suggested that across adolescence there are general decreases in antisocial behaviour, ADHD symptoms, anxiety, depression and prosociality, with a possible rebound in late adolescence for the latter. However, these average trends are in the context of considerable variation across individuals with, for example, growth mixture analyses, often revealing a non-trivial subgroup for whom levels of these dimensions increase across development (e.g. Crocetti et al., 2009). These dimensions may influence one another through developmental cascades. For example, early conduct problems may put an individual at risk of academic and social difficulties that in turn increase the risk of anxiety and depression (e.g. Van Lier et al., 2012).

The above-mentioned studies rely on trait estimates that can be validly compared across development; however, The array and pace of changes that occur during adolescence makes this a challenge.. Over development, constructs may change in nature; particular indicators maylose their developmental appropriateness ; or samples may reach a floor or ceiling on items that previously discriminated well between different trait levels (e.g. Edwards & Wirth, 2012). Early adolescence, for example, sees a spike in certain types of anti-social behaviour such as rule-breaking and non-compliance with authority. For most individuals, these behaviours decline into late adolescence and adulthood (e.g. Barker et al., 2007). As such, endorsing an item measuring one of these behaviours would tend to index a greater degree of underlying severity in late adolescence than the same behaviour endorsed in early adolescence. Similarly, the attention deficit versus hyperactivity/impulsivity symptoms of ADHD have previously shown evidence of differential developmental trajectories. The former has shown a potential curvilinear trajectory while the lattershow a more definite and steady decline over development (e.g. Murray, Obsuth et al., 2016). Over the course of adolescence, this could, for example, lead to increasing bifurcation of a previously unitary

ADHD construct into separate inattention and hyperactivity/impulsivity constructs. More generally, items referring to symptoms or behaviours within the context of typical early adolescent social environments and activities (e.g. school) may be less relevant by late adolescence.

When items cannot be considered comparable across time, valid inferences about change in the underlying trait rely on identifying and modelling the nature and extent of non-comparability. 'Comparability' comes in different degrees and types that can be modelled and tested statistically. A useful framework for doing so is that of longitudinal measurement invariance within confirmatory factor analysis (CFA) (e.g. see Millsap & Cham, 2012). Longitudinal measurement invariance is when expected observed score distributions given trait levels are independent of the wave at which the measure was administered. Longitudinal invariance using CFA tests a slightly weaker version of this; namely, that the expected mean and variance of observed score distributions given latent trait levels are independent of measurement wave. In this framework, a measure may show (from weakest to strongest): no invariance, configural invariance, metric invariance, scalar invariance or residual invariance.

Configural variance means that only the pattern of factor loadings for a measure is the same across time. Metric invariance is where both factor loading patterns and factor loading magnitudes are equal across time. When metric invariance holds, comparisons of factor variances and covariances are supported, for example, in cross-lagged and autoregressive panel models. When metric invariance does not hold, an inventory may not be measuring the same construct across time. Scalar invariance is when factor patterns loadings and intercepts are equal across time. When scalar invariance holds, inferences about mean differences over time are supported. This is relevant when, for example, fitting second-order growth curve models. Finally, residual (or strict) invariance is when factor patterns, loadings, intercepts and item residual variances are equal over time. When this holds, differences in means and

variances of the observed scores can be attributed to differences on the underlying latent factors. Only in this case are inferences about stability and change in a trait over time based on observed scores (e.g. sum scores) supported. Where any level of the above-described invariance assumptions are not met, it is often possible to nonetheless obtain valid comparisons of latent constructs across time. To do so requires that at least some items are longitudinally invariant, giving 'partial invariance'. Partial invarianceoften suffices provided that the non-invariance in the remaining items is explicitly modelled (e.g. Edwards & Wirth, 2012).

Surprisingly few studies of adolescent development report longitudinal invariance analyses either in their own right or to support other longitudinal analyses. Some studies have suggested that high levels of invariance across adolescence can be achieved for specific measures of mental health dimensions health such as ADHD, depression, anxiety and conduct problems (e.g. Leopold et al., 2016; Motl et al., 2005; Sterba et al., 2010; Verhoeven et al., 2013) while others have identified non-invariance (Mathyssek et al., 2013; Sterba et al., 2010). In this study, we sought to build on this thus far limited evidence base and evaluate longitudinal invariance in brief measures of 5 dimensions of mental health across development. We aimed to evaluate the kind of inferences that are supported in longitudinal analyses using the self-reported Social Behavior Questionnaire (SBQ; Tremblay et al., 1991) and whether there is any risk of bias when using sum scores. We also aimed to establish what longitudinal measurement models are likely to be needed to obtain comparable estimates across development. Although versions of the measure are used in several large child and adolescent development studies internationally, there is as yet no evidence of its longitudinal measurement properties (e.g. Lösel and Stemmler, 2012; Rouquette, et al., 2014).

The Social Behavior Questionnaire shares origins with the Strengths and Difficulties Questionnaire (SDQ; Goodman, 1997). Its first reported use was in a study by Tremblay et al.

(1991) where two pre-existing scales were combined: 28 items from the Preschool Behavior Questionnaire (Behar & Stringfield, 1974), itself an adaptation of the Children's Behavior Questionnaire (Rutter, 1967) and 10 items from the Prosocial Behavior Questionnaire (Weir & Duveen, 1981). Though originally developed for young children, it has since been widely applied in older children and adolescents. It has been adopted in several large-scale studies internationally, where variants have been administered in English, French, and German. For example, items of the SBQ were integrated in the behaviour evaluations of adopted in the Canadian National Longitudinal Survey of Children and Youth (Sprott, 2000); the Nuremberg-Erlangen Prevention and Development Study (e.g. Lösel & Stemmler, 2012) and The Quebec Longitudinal Study of Kindergarten Children (QLSKC; Rouquette, et al., 2014). It is used in self-, parent- and teacher- report form.

In spite of the large number of published studies using the SBQ items, there have been only a small number of previous dedicated studies of its psychometric properties. The few that have been conducted have generally supported the factorial validity, criterion validity and reliability of the SBQ items (e.g. Murray, Eisner & Ribeaud, 2017; Murray, Eisner & Ribeuad, 2016b; Murray, Eisner & Ribeaud, 2017; Tremblay et al., 1991; Tremblay et al., 1992). Murray, Eisner and Ribeaud (2017), for example, analysed the ranges of measurement for which the teacher-reported SBQ items reliably captured the dimensions of prosociality, externalising, ADHD, and internalising in z-proso at ages 7 through to 15. The latter three traits were operationalised using bi-factor measurement models where the specific dimensions of ADHD were inattention and hyperactivity/impulsivity, the specific dimensions of internalising were anxiety and depression, and the specific dimensions of externalising were physical aggression, oppositional defiant disorder/conduct disorder, and reactive aggression. They found that SBQ items could reliably capture a wide range of trait levels for the general factors across ages 7 to 15. Murray, Obsuth et al. (2017) reported a factor analysis

of all self-reported SBQ items together with the newly developed *Violent Ideations Scale* (Murray, Eisner & Ribeaud, 2016b) in the z-proso sample when the participants were aged 17. All items from the SBQ loaded on the intended factors (prosociality, internalising, ADHD, reactive aggression and proactive aggression). In contrast to the SBQ reliability study by Murray, Eisner & Ribeaud (2017), this study did not support a distinction between anxiety and depression; however, Murray, Obsuth et al. (2017) cited concerns about over-factoring and it is possible that extracting further factors would have yielded separate anxiety and depression factors. Arguably, understanding the longitudinal measurement properties of an instrument is an important step in its validation and in construct validation more broadly (e.g. Edwards & Wirth, 2009); however, this has not been tested for any version of the SBQ across any developmental period. A second goal of our study was to explore whether any non-invariance identified could provide insights into the manner in which self-perception, references frames, levels of disclosure or the manifestation of constructs change over adolescent development.

Method

Participants

Data came from the Zurich Project on the Social Development of Children and Youths (z-proso). This is a longitudinal cohort study of youth, focussing on the development of prosocial and anti-social behaviours. The study began in 2004 when the participants were aged 7 and entering school. Participants were selected according to a school-level stratified random sampling procedure that took into account school size and location. All children entering the first grade at selected schools in that year were invited to participate. The study

included separate child and parent intervention components in the early waves; however, as there was little evidence that they had any substantive short or long term effects it is usually judged reasonable to treat the data as observational (e.g. Averdijk, Zirk-Sadowski, Ribeaud & Eisner, 2016; Malti, Ribeaud & Eisner, 2011). Comprehensive accounts of recruitment, participant characteristics, attrition and assessment procedures can be found in previous publications (e.g. Eisner & Ribeaud, 2007) and on the z-proso website: http://www.cru.ethz.ch/en/projects/z-proso.html. The current study focusses on the latter 4 measurement waves when the majority of participants were aged 11, 13, 15 and 17 respectively. Across these waves 1523 (51% male) participants contributed data, representing 91% of the original target sample.

Measures

Prosociality, Anxiety, Depression, ADHD and Aggression

Prosociality, anxiety, depression, ADHD and Aggression were measured with the Social Behavior Questionnaire (SBQ).

Z-proso includes some adaptations of the SBQ, mainly additional aggression items reflecting the fact that a major research theme of the study is antisocial behaviour development. In addition, whereas the original SBQ was on a three-point response scale, the version administered in z-proso offers respondents a five-point scale. In our longitudinal invariance analyses, we focus on items that were common across all measurement waves to simplify the interpretation of any non-invariance identified. The majority of these items were part of the original SBQ reported in Tremblay et al., (1991). One anxiety item, two depression items, two aggression items were added in z-proso to maintain developmental appropriateness in the adolescent period. Overall, the SBQ administered in z-proso included 8

items measuring prosociality that were common across all measurement waves (age 11, 13, 15, and 17); 4 measuring depression; and 4 measuring anxiety. In addition, 4 items measuring ADHD and 8 measuring aggression (4 each for the subtypes of proactive and reactive aggression) were administered at the latter three measurement waves (age 13, 15 and 17) only. Thus, we analyse a slightly different set of items as compared to the original SBQ items developed for preschoolers. Abbreviated item contents are provided in the results tables. All items are measured on a 5 point Likert scale from *never* to *very often.* Apart from the anxiety and depression items which referred to the frequency of behaviour/symptoms in the past month, all items referred to frequency of behaviour/symptoms in the past year. All were administered in paper and pencil format in German: the official language of the study location.

Statistical procedure

Longitudinal factorial invariance

Longitudinal factorial invariance was assessed within a confirmatory factor analysis (CFA) framework. We treated items as continuous because it is usually reasonable to treat ordered-categorical items with at least five response options as continuous (e.g. Rhemtulla, Brosseau-Laird & Savalei, 2012). Benefits of doing so include better accounting for missingness (because FIML can be employed), a larger simulation evidence base to draw on to guide model selection based on model fits (e.g. Chen, 2007; Cheung & Rensvold, 2002; Meade et al., 2008), and the availability of information theoretic criteria to further guide model selection (e.g. Raftery, 1995). Further, Sass et al. (2014) presented simulation study results suggesting that fit comparisons for invariance testing that treats items as ordered categorical using weighted least squares means and variances (WLSMV) do not perform well. Comparisons across a series of increasingly constrained models provided information on the level of invariance that could be achieved for each trait.. We report e χ^2difference

tests for information but note that these are likely to be overly sensitive to minor mis-specifications given our large sample size (e.g. Meade et al., 2008). As such, for model selection we relied primarily on the approximate fit indexes of CFI, TLI and RMSEA which are less influenced by sample size and model complexity than the χ^2difference test. In particular, we use the cut-off criteria defined by Chen (2007), developed using a simulation study of different types of invariance at different sample sizes. According to these criteria, metric invariance would hold in our sample when the CFI decreased by less than 0.010, RMSEA increased by less than 0.015 and SRMR increased by less than 0.030 with the addition of metric invariance constraints. Scalar invariance would hold if CFI decreased by less than 0.010, RMSEA increased by less than 0.015 and SRMR increased by less than 0.010 with the addition of scalar invariance constraints. Residual (or strict) invariance would hold if CFI decreased by less than 0.010, RMSEA increased by less than 0.015 and SRMR increased by less than 0.010 with the addition of scalar invariance constraints. Noting that these criteria did not have power to detect minor violations of invariance, Meade et al. (2008) suggest that a change in CFI of >.002 to indicate non-invariance; however, given the large number of comparisons to be conducted overall, we elected to use Chen's (2007) more conservative criteria to avoid the detection of a large numbers of instances of trivially small non-invariance.

To test configural, metric, scalar and residual invariance, a series of increasingly strict models were fit to test invariance. First, in the configural model, patterns of loadings were fixed equal over time but loadings and intercepts were free to vary. Scaling and identification were achieved by fixing the mean and variance of the latent factor at baseline to 0 and 1 respectively and by fixing the loading and threshold of a reference item to equality over all time points. This method assumes that the reference item is invariant over time. Choosing a reference item that is not invariant can lead to the appearance of non-invariance

in truly invariant items and/or to the appearance of invariance in truly non-invariant items (e.g. Yoon & Millsap, 2007). However, there was no prior empirical evidence nor strong theoretical rationale on which to base the selection of reference variable for the SBQ. As such, we provisionally selected the first item in each subscale to be the reference variable and then checked that the invariance constraints with these items were not associated with large modification indices at either the metric or scalar model stage. If they were, we allowed these constraints to be freed and instead thereafter relied on items with smaller modification indices associated with their invariance constraints to act as reference variables.

Residual covariances between the same item measured over time were also estimated. Configural invariance was judged to hold when the configural model had RMSEA<.08, SRMR<.08 and TLI and CFI >.95 (e.g. Hu & Bentler, 1999; Schermelleh-Engel et al., 2003).In the metric model, equality constraints over time were added on the remaining first-order loadings.

In the scalar model, equality constraints over time were added on the remaining intercepts.

In the strict invariance model, residual variances were fixed equal over time.

If at any stage invariance was not supported, modification indices and expected parameter changes were used to identify specific constraints that did not hold. These were iteratively released, re-testing invariance with the release of each individual constraint until either partial invariance held or there were only two items left with invariance constraints at that level. In this latter case, invariance was judged not to hold at that level. No further constraints were added to items that did not show invariance at a lower level e.g. scalar constraints were not placed on items that did not show metric invariance.

Finally, the practical consequences of non-invariance were investigated. To do this, we compared two estimates. The first was a linear slope factor mean from a growth curve models fit the data using a model that assumed full invariance (all loadings, intercepts and residual variance assumed equal over time) versus a model that assumes the highest level of invariance that was actually attained. The difference provides a quantification of the bias that could be expected to result from the non-invariance, if not appropriately modelled.

Results

Overview

Invariance was not supported at any level according to the χ^2difference test; however, as discussed above this test is likely not appropriate for our large sample size as it will identify even trivially small mis-specifications as statistically significant. Partially invariant models were achieved in all cases according to the fit criteria of Chen (2007). The practical significance of the non-invariance identified was overall small: unstandardised slope estimates from partially invariant versus (falsely assumed) fully invariant models generally differed only at the second decimal place. The one exception was anxiety, for which slopes were overestimated by more than 50% when longitudinal invariance was incorrectly assumed.

Prosociality

Fits for the prosociality models are provided in Table 1. The configural model was a single factor model. Both configural and metric invariance held but scalar invariance did not. Iterative release of equality constraints on intercepts, in the following sequence resulted in a partially invariant model (M2a in Table 2): item 48 at age 11; item 47 at age 17; item 41 at age 17; and item 48 at age 13. Adding residual invariance constraints, fit statistics suggested further non-invariance. Based on modification indices and expected parameter changes,

constraints on the item residual variances for item 47 at age 11 and then item 41 at age 11 were released. At this point, partial residual invariance was judged to hold. Parameters for this final model (M3a) are provided in Table 2. These show that for item 48, the intercept increased over time. In addition, the residual variances for items 41 and 47 were larger at age 11 than at subsequent measurement points. The linear slope factor mean from a latent growth curve model fit using the final measurement model developed in these analyses was -0.290. The model assuming full measurement invariance across development yielded a corresponding estimate of -0.209.

Anxiety

Fits for the anxiety models are provided in Table 3. The configural model was a single factor model. Both configural and metric invariance held. Scalar invariance did not hold but partial invariance was achieved with the iterative release of scalar constraints on item 1 measured at age 11 and then the same item measured at age 13. No further non-invariance was identified with the addition of residual invariance constraints on the scalar invariant items. Partial residual invariance was, therefore, judged to hold. Parameters for this model are provided in Table 4. These show that the intercept for item 4 increased across ages 11, 13 and 15. The linear slope factor mean from a latent growth curve model fit using the final measurement model developed in these analyses was 0.257. The model assuming full measurement invariance across development yielded a corresponding estimate of 0.544.

Depression

Fits for the depression models are provided in Table 5. The configural model was a single factor model and showed good fit. Both configural and metric invariance held but scalar invariance did not. The scalar invariance constraint on item 63 at age 13 was removed to achieve partial scalar invariance. Partial residual invariance was achieved with the

addition of residual invariance constraints to all remaining items. The parameters from this model are provided in Table 6. Item 4 had a larger intercept at age 13. The linear slope factor mean from a latent growth curve model fit using the final measurement model developed in these analyses was 0.716. The model assuming full measurement invariance across development yielded a corresponding estimate of 0.725.

ADHD

Fits for the ADHD models are provided in Table 7. The configural model for ADHD was a single-factor model. Configural, metric, scalar and residual invariance all held. Parameter estimates from the residual invariance model are provided in Table 8. The slope factor mean for a linear growth curve model using this fully invariant ADHD model was 0.114.

Aggression

Fits for the aggression models are provided in Table 9. The configural model was a first-order oblique model with first-order factors: proactive aggression and reactive aggression. The configural model showed reasonable fit and configural invariance was judged to hold. Metric but not scalar invariance held. Release of the intercept constraint on item 61 at age 17 was necessary to achieve partial scalar invariance. To achieve partial residual invariance it was necessary to release the constraints on the residual variance of item 37 at age 17 and then on item 72 at age 13. Parameter estimates for this model are provided in Table 10. These show that the intercept for item 61 was lower at age 17 while the residual variance for items 37 and 72 was larger at earlier waves. The linear slope factor mean from a latent growth curve model fit using the final measurement model developed for reactive aggression was -0.041 and for proactive aggression was -0.210. The models assuming full

measurement invariance across development for these constructs yielded corresponding estimates of -0.084 for reactive aggression and -0.202 for proactive aggression.

Discussion

In the current study, we examined the important but seldom-tested assumption of longitudinal factorial invariance for five dimensions of adolescent mental health measured by the Social Behavior Questionnaire (SBQ). We found that items largely functioned equivalently across waves, supporting their use in longitudinal analyses. Where non-invariance was identified, the most consistent pattern was larger residual variances at earlier time points. This suggests that the adolescents may have become better able to report on the presence of certain behaviours or symptoms over time. The potential bias arising from falsely assuming invariance with these data (e.g. by using sum scores) is not likely to be substantial , except for developmental analyses involving anxiety.

Longitudinal invariance analyses are rarely reported in studies of adolescent development. This is in spite of the fact that the majority of conclusions drawn about psyschosocial development from longitudinal data rely on the assumption that at least metric invariance holds. Further, the selection of measures for longitudinal studies seldom explicitly considers the degree of comparability of measures over the relevant phases of development as a criterion. In this study, we this evaluated longitudinal invariance for five dimensions of adolescent mental health measured by the Social Behavior Questionnaire (SBQ). All subscales showed at least metric invariance which is consistent with the idea that they measure the same constructs over adolescence. This makes the SBQ a good candidate for use as a brief omnibus measure of mental health dimensions in future studies of adolescent psychosocial development.

Scalar invariance was violated for some items across the SBQ subscales. Here there was a general tendency for item intercepts to increase over measurement waves. As such, for a given latent trait level, expected item scores were higher at later time points. The implication of this is that a researcher using observed scores (e.g. sum score for a subscale) would see an artefactual increase (or an attenuated decrease) in levels of the traits measured with these items over adolescence. However, as so few items were affected, it is likely that unbiased comparisons of levels over time could be made using a partial measurement invariance model in which the scalar invariance constraints that did not hold were not imposed (Byrne et al., 1989; also see Ferrer et al., 2008). Indeed, comparisons of latent growth curve models from models appropriately modelling non-invariance versus models that assumed full invariance revealed substantial discrepancies only for anxiety. For anxiety, the positive slope over time was overestimated by 52% because the model mis-attributed an increase in an item intercept to an increase in factor means across time. A possible explanation for the trend towards increasing intercepts over time is that adolescents become more attuned to the behaviours and symptoms about which the items ask over time. This could lead to a trend towards being more likely to endorse these items over time. This could in turn be a function of the repeated administration of the questionnaires whereby participants are cumulatively primed to detect certain symptoms and behaviours. It could also be a function of increasing capacity to to identify and report on these same symptoms and behaviours arising from maturity. An alternative explanation would be that participants become more comfortable disclosing information about their negative behaviours and symptoms over time as they build trust with the study over measurement waves; however, the fact that the same trend was seen in prosociality - a trait with positive connotations - calls this interpretation into question.. The one item that deviated from this pattern was an item measuring a tendency to experience boredom, an indicator of depression. This item had a

larger intercept at age 13 than at other ages. One possibility is that this reflects a normative elevation of boredom specifically around this age (Spaeth et al., 2015).

There were also some violations of residual invariance. In general, residual variances were larger at age 11, suggesting that measurement error decreased over time. This is likely to reflect increased familiarity with the questionnaire on repeated administrations and an increase in the reliability of self-reports that comes with maturity.. However, the fact that these violations were few and generally of a small magnitude suggests that by age 11, self-reports are not substantially less reliable than those at older ages.

There are implications of these findings both for users of the z-proso dataset specifically and researchers of adolescent development more generally. Users of the dataset should ideally adopt the measurement models for the anxiety, depression, prosociality and aggression outlined in the Results section, especially for anxiety. For certain models with a high degree of computational complexity - such as those involving a large number of latent interactions- using factor scores estimated from the measurement models would represent a more practical solution. In these cases, the researcher should confirm that the determinacy of the factor scores is adequate, e.g. >.90 (Gorsuch, 1983). For ADHD – which showed full invariance over development – using sum scores over latent variable measurement models would not be expected to introduce substantial bias. However, there would nonetheless be other benefits to using a latent variable measurement model such as greater power due to the disattenuation for unreliability and the ability to test model fit and identify mis-specifications.

While these results provide validity support for the SBQ in showing that most items remain comparable across development, users of the SBQ in other studies should independently investigate invariance and develop appropriate measurement models to take account of any violations identified. There are no guarantees that invariance results would

generalise across studies because the methodological differences between studies using the SBQ and the different languages and cultural contexts in which it is administered could influence longitudinal invariance. For example, the SBQ is administered on a three- rather than five- point response format in some studies. Comparisons of invariance results across different studies using the SBQ would help to home in on the features generating the non-invariance.

More generally, our study illustrates that in spite of the myriad changes occurring across adolescence, it is possible to construct measures that capture dimensions such as anxiety, depression, ADHD, aggression and prosociality in a comparable way across development. This is critical to robust research into developmental trends in these dimensions. Any study seeking to make inferences about changes in variances, covariances and means (or statistics derived from these) should follow a procedure similar to that outlined in our Method and Result section, in order to build appropriate measurement models to account for non-invariance prior to testing their substantive hypotheses. In other cases, it may be necessary to also investigate invariance with respect toother variables at the same time. For example, tests differential subgroup developmental trajectories (e.g. sex differences in development) would call for investigations of invariance by both subgroup and measurement wave.

The fact that the SBQ showed high levels of invariance over time likely reflects the fact that the wording of items is relatively general and not tied to specific contexts (e.g., school) or activities (e.g. schoolwork). When planning longitudinal studies, including a core set of 'context-free' or 'developmentally neutral' items that could be expected to show invariance over development may be crucial to ultimately obtaining comparable trait estimates over longer time spans. This is not to discourage the inclusion of items that are specific to some developmental periods; omitting these may result in crucial manifestations

of traits being missed. For these items to be included in studies of change over time; however, they need to be anchored via developmentally invariant items (e.g., Edwards & Wirth, 2009). The SBQ items may, for example, serve as useful anchors when developing new measures for longitudinal studies or administered alongside more comprehensive measures of the dimensions it measures.

Finally, it is important to consider the limitations of the approach of the current paper. First, factorial invariance does not guarantee that a measure captures the same thing over time; it is possible for metric invariance to hold, for example, when the psychological process underlying responses is changing (e.g. Widaman et al., 1992). Second, we used relatively conservative fit criteria for detecting invariance to protect against making Type 1 errors. However, this also increases the likelihood of missing minor violations of invariance. Third we focused on items that were identical over multiple waves. As such, the measures may not have been unrepresentative of the traits during particular time periods. Fourth, although overall tests of invariance at the scale level tend to be relatively robust, this is not necessarily true at the item level where, for example, iterative release of constraints based on modification indices or the selection of a non-invariant reference variable can risk mis-identifying the location of non-invariance (e.g. Johnson, Meade & DuVernet, 2009). As such, greater caution is due regarding item-level inferences regarding invariance. Finally, we did not have any a priori hypotheses regarding which items should be non-invariant and in what way (e.g. whether intercepts should increase over development). As such, our interpretations of the non-invariance were entirely post hoc. Future studies may be able to derive a priori hypotheses about non-invariance across time from developmental theories of the traits (s) being analysed.

Conclusions

Although adolescence is time of considerable change, it is possible to identify items that show comparable measurement properties across this period, a prerequisite for supporting valid inferences about development. Where measurement invariance is violated, it tends to be the case that residual variances are larger at earlier waves while intercepts are larger at later waves. In the SBQ, only the anxiety subscale showed practically significant levels of non-invariance.

Table 1: Model fits for Prosociality invariance models

Model	Model Description	χ^2	Df	*P*	CFI	TLI	RMSEA	SRMR	AIC	BIC
					Model fits					
M0	Configural	885.052	410	<.001	0.965	0.958	0.028	0.033	106087.155	106886.421
M1	Metric	926.658	431	<.001	0.963	0.958	0.027	0.036	106086.761	106774.129
M2	Scalar	1556.853	452	<.001	0.918	0.910	0.040	0.043	106674.956	107250.428
M2a	Scalar partial	1089.311	448	<.001	0.953	0.948	0.031	0.037	106215.415	106812.200
M3	Strict (partial)	1255.434	468	<.001	0.942	0.938	0.033	0.054	106341.538	106831.754
M3a	**Strict partial**	**1168.715**	**466**	**<.001**	**0.948**	**0.945**	**0.031**	**0.047**	**106258.818**	**106759.691**
					Model fit differences					
M1-M0	Metric-configural	41.606	21	.005	-0.002	0	-0.001	0.003	-0.394	-112.292
M2-M1	Scalar-Metric	630.195	21	<.001	-0.045	-0.048	0.013	0.007	588.195	476.299
M2a-M1	Scalar partial- Metric	162.653	17	<.001	-0.01	-0.01	0.004	0.001	128.654	38.071
M3-M2a	Strict (partial)- Scalar partial	166.123	20	<.001	-0.011	-0.01	0.002	0.017	126.123	19.554
M3a-M2a	Strict partial- Scalar partial	79.404	18	<.001	-0.005	-0.003	0	0.01	43.403	-52.509

Note. Final model indicated in bold.

Table 2: Parameters for most invariant prosociality model

	Loadings				**Intercepts**				**Residual variances**			
Item number and content	**Age 11**	**Age 13**	**Age 15**	**Age 17**	**Age 11**	**Age 13**	**Age 15**	**Age 17**	**Age 11**	**Age 13**	**Age 15**	**Age 17**
41- help clear up	0.423	0.423	0.423	0.423	3.335	3.335	3.335	**3.558**	**0.987**	0.793	0.793	0.702
48- understand feelings	0.515	0.515	0.515	0.515	**3.621**	**3.917**	4.143	4.143	0.888	0.741	0.563	0.563
49- share	0.445	0.445	0.445	0.445	3.855	3.855	3.855	3.855	0.696	0.696	0.696	0.696
42- settle dispute	0.627	0.627	0.627	0.627	3.413	3.413	3.413	3.413	0.796	0.796	0.796	0.796

40- sympathy	0.719	0.719	0.719	0.719	4.016	4.016	4.016	4.016	0.487	0.487	0.487	0.487
43- help injured	0.696	0.696	0.696	0.696	4.042	4.042	4.042	4.042	0.470	0.470	0.470	0.470
46- comfort	0.841	0.841	0.841	0.841	4.032	4.032	4.032	4.032	0.415	0.415	0.415	0.415
47- listen to others	0.472	0.472	0.472	0.472	3.615	3.615	3.615	**3.930**	**1.211**	0.800	0.800	0.649
Factor means	Age 11= 0, Age 13= -0.368, Age 15=-0.336, Age 17= -0.242											
Factor variances	Age 11= 1, Age 13= 1.086, Age 15=0.923, Age 17= 0.919											

Note. Non-invariant parameters are indicated in bold.

Table 3: Model fits for Anxiety invariance models

Model	Model description	χ^2	*df*	*p*	CFI	TLI	RMSEA	SRMR	AIC	BIC
Model fits										
M0	Configural	212.222	74	<.001	.979	.966	.035	.029	57861.782	58277.4
M1	Metric	247.734	83	<.001	.975	.963	.036	.034	57879.294	58246.956
M2	Scalar	541.796	92	<.001	.931	.91	.057	.057	58155.355	58475.061
M2a	Scalar partial	278.714	90	<.001	.971	.961	.037	.036	57896.273	58226.636
M3	**Strict (partial)**	**306.053**	**100**	**<.001**	**.968**	**.962**	**.037**	**.044**	**57903.613**	**58180.692**
Model fit differences										
M1-M0	Metric-configural	35.512	9	<.001	–.004	–.003	.001	.005	17.512	–30.444
M2-M1	Scalar-metric	148.807	9	<.001	–.044	–.053	.021	.023	276.061	228.105
M2a-M1	Scalar partial-metric	30.98	7	<.001	–.004	–.002	.001	.002	16.979	–20.32
M3-M2a	Strict (partial)-scalar partial	27.339	10	.002	–.003	.001	0	.008	7.34	–45.944

Note. Final model indicated in bold.

Table 4: Parameters for most invariant anxiety model

	Loadings				**Intercepts**				**Residual variances**			
Item number	**Age 11**	**Age 13**	**Age 15**	**Age 17**	**Age 11**	**Age 13**	**Age 15**	**Age 17**	**Age 11**	**Age 13**	**Age 15**	**Age 17**
01- crying	0.662	0.662	0.662	0.662	2.034	2.034	2.034	2.034	0.573	0.573	0.573	0.573
03- fear	0.574	0.574	0.574	0.574	1.637	1.637	1.637	1.637	0.478	0.478	0.478	0.478
04- worry	0.668	0.668	0.668	0.668	**2.070**	**2.322**	2.637	2.637	0.748	0.793	0.815	0.815
62- sleepless	0.504	0.504	0.504	0.504	2.463	2.463	2.463	2.463	1.274	1.274	1.274	1.274
Factor means	Age 11=0, Age 13=0.052, Age 15=0.162, Age 17= 0.276											
Factor variances	Age 11=1, Age 13=1.207, Age 15=1.494, Age 17=1.699											

Note. Non-invariant parameters are indicated in bold.

Table 5: Model fits for depression invariance models

Model	Model description	χ^2	*df*	*p*	CFI	TLI	RMSEA	SRMR	AIC	BIC
Model fits										
M0	Configural	109.904	74	.0043	.994	.991	.018	.022	56587.043	57002.661
M1	Metric	139.798	83	<.001	.991	.987	.021	.027	56598.937	56966.599
M2	Scalar	276.202	92	<.001	.971	.962	.036	.035	56717.341	57037.047
M2a	Scalar partial	197.280	91	<.001	.983	.978	.028	.029	56640.419	56965.454
M3	**Strict (partial)**	**245.175**	**102**	**<.001**	**.977**	**.974**	**.030**	**.035**	**56666.314**	**56932.736**
Model fit differences										
M1-M0	Metric-configural	29.894	9	<.001	–.003	–.004	.003	.005	11.894	–36.062
M2-M1	Scalar-metric	136.404	9	<.001	–.02	–.025	.015	.008	118.404	70.448
M2a-M1	Scalar partial-metric	57.482	8	<.001	–.008	–.009	.007	.002	41.482	–1.145
M3-M2a	Strict (partial)-scalar partial	47.895	11	<.001	–.006	–.004	.002	.006	25.895	–32.718

Note. Final model indicated in bold.

Table 6: Parameters for most invariant depression model

	Loadings				**Intercepts**				**Residual variances**			
Item number	**Age 11**	**Age 13**	**Age 15**	**Age 17**	**Age 11**	**Age 13**	**Age 15**	**Age 17**	**Age 11**	**Age 13**	**Age 15**	**Age 17**
05- sad no reason	0.595	0.595	0.595	0.595	1.731	1.731	1.731	1.731	0.829	0.829	0.829	0.829
08- unhappy	0.68	0.68	0.68	0.68	2.051	2.051	2.051	2.051	0.516	0.516	0.516	0.516
63- bored	0.269	0.269	0.269	0.269	2.634	**2.870**	2.634	2.634	0.833	0.867	0.833	0.833
64- feel alone	0.721	0.721	0.721	0.721	1.704	1.704	1.704	1.704	0.468	0.468	0.468	0.468
Factor means	Age 11=0, Age 13=0.223, Age 15=0.575, Age 17= 0.728											
Factor variances	Age 11=1, Age 13=1.337, Age 15=1.657, Age 17=1.827											

Note. Non-invariant parameters are indicated in bold.

Table 7: Model fits for ADHD invariance models

Model	Model description	χ^2	*df*	*p*	CFI	TLI	RMSEA	SRMR	AIC	BIC
Model fits										
M0	Configural	187.202	39	<.001	.969	.948	.051	.029	43308.221	43578.614
M1	Metric	215.756	45	<.001	.964	.948	.051	.036	43324.775	43563.357
M2	Scalar	251.771	51	<.001	.958	.946	.052	.037	43348.790	43555.561
M3	**Strict**	**302.348**	**59**	**<.001**	**.949**	**.943**	**.053**	**.045**	**43383.367**	**43547.724**
Model fit differences										
M1-M0	Metric-configural	28.554	6	<.001	−.005	0	0	.007	16.554	−15.257
M2-M1	Scalar-metric	36.015	6	<.001	−.006	−.002	.001	.001	24.015	−7.796
M3-M2	Strict-scalar	50.577	8	<.001	−.009	−.003	.001	.008	34.577	−7.837

Note. Final model indicated in bold.

Table 8: Parameters for most invariant ADHD model

	Loadings			**Intercepts**			**Residual variances**		
Item number	**Age 13**	**Age 15**	**Age 17**	**Age 13**	**Age 15**	**Age 17**	**Age 13**	**Age 15**	**Age 17**
12- restless	0.729	0.729	0.729	2.761	2.761	2.761	0.582	0.582	0.582
13- concentration	0.726	0.726	0.726	2.348	2.348	2.348	0.512	0.512	0.512
16- inattention	0.652	0.652	0.652	2.684	2.684	2.684	0.67	0.67	0.67
17- hectic/fidgety	0.539	0.539	0.539	2.741	2.741	2.741	0.739	0.739	0.739
Factor means	Age 13=0, Age 15=0.105, Age 17=0.304								
Factor variances	Age 13=1, Age 15=1.046, Age 17=1.126								

Note. Non-invariant parameters are indicated in bold.

Table 9: Model fits for aggression invariance models

Model	Model description	χ^2	*df*	*p*	CFI	TLI	RMSEA	SRMR	AIC	BIC
Model fits										
M0	Configural	1161.785	213	<.001	.919	.895	.055	.054	70507.736	71096.239
M1	Metric	1189.546	225	<.001	.917	.899	.054	.055	70511.498	71036.378
M2	Scalar	1400.582	237	<.001	.900	.884	.058	.056	70698.534	71159.792
M2a	Scalar partial	1273.555	236	<.001	.911	.896	.054	.055	70573.506	71040.066
M3	Strict (partial)	1464.645	251	<.001	.896	.886	.057	.060	70734.597	71121.630
M3a	**Strict partial**	**1382.150**	**249**	**<.001**	**.903**	**.892**	**.055**	**.057**	**70656.101**	**71053.738**
Model fit differences										
M1-M0	Metric-configural	27.761	12	.006	–.002	.004	–.001	.001	3.762	–59.861
M2-M1	Scalar-metric	211.036	12	<.001	–.017	–.015	.004	.001	187.036	123.414
M2a-M1	Scalar partial-metric	84.009	11	<.001	–.006	–.003	0	0	62.008	3.688
M3-M2a	Strict (partial)-scalar partial	191.09	15	<.001	–.015	–.01	.003	.005	161.091	81.564
M3a-M2a	Strict partial-scalar partial	108.595	13	<.001	–.008	–.004	.001	.002	82.595	13.672

Note. Final model indicated in bold.

Table 10: Parameters for most invariant aggression model

	Loadings			**Intercepts**			**Residual variances**		
Item number	**Age 13**	**Age 15**	**Age 17**	**Age 13**	**Age 15**	**Age 17**	**Age 13**	**Age 15**	**Age 17**
Proactive aggression									
37- threaten	0.368	0.368	0.368	1.235	1.235	1.235	0.230	0.230	**0.156**
51- bossing	0.503	0.503	0.503	1.740	1.740	1.740	0.500	0.500	0.500
52- forcing	0.688	0.688	0.688	1.579	1.579	1.579	0.295	0.295	0.295
61- humiliate	0.632	0.632	0.632	1.891	1.891	**1.621**	0.551	0.551	**0.373**
Reactive aggression									
53- when teased	0.407	0.407	0.407	3.000	3.000	3.000	1.033	1.033	1.033
54- when something taken	0.657	0.657	0.657	1.620	1.620	1.620	0.295	0.295	0.295

55- when not getting something	0.374	0.374	0.374	1.797	1.797	1.797	0.630	0.630	0.630
72- when insulted	0.766	0.766	0.766	1.705	1.705	1.705	**0.335**	0.203	0.203
Factor means	Proactive aggression: Age 13=0, Age 15=-0.05, Age 17 = -0.104 Reactive aggression: Age 13=0, Age 15=-0.213, Age 17= -0.445								
Factor variances	Proactive aggression: Age 13=1, Age 15=0.937, Age 17=0.742 Reactive aggression: Age 13= 1, Age 15=0.832, Age 17=0.742.								

Note. Non-invariant parameters are indicated in bold.

References

Averdijk, M., Zirk-Sadowski, J., Ribeaud, D., & Eisner, M. (2016). Long-term effects of two childhood psychosocial interventions on adolescent delinquency, substance use, and antisocial behavior: a cluster randomized controlled trial. *Journal of Experimental Criminology*, *12,* 21-47.

Barker, E. D., Séguin, J. R., White, H. R., Bates, M. E., Lacourse, E., Carbonneau, R., & Tremblay, R. E. (2007). Developmental trajectories of male physical violence and theft: relations to neurocognitive performance. *Archives of General Psychiatry*, *64*, 592-599.

Behar, L., & Stringfield, S. (1974). *Manual for the Preschool Behavior Questionnaire.* Lenore Behar: Durham, NC.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, *105*, 456-466.

Carlo, G., Crockett, L. J., Randall, B. A., & Roesch, S. C. (2007). A latent growth curve analysis of prosocial behavior among rural adolescents. *Journal of Research on Adolescence*, *17*, 301-324.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling*, *14*, 464-504.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*, 233-255.

Crocetti, E., Klimstra, T., Keijsers, L., Hale III, W. W., & Meeus, W. (2009). Anxiety trajectories and identity development in adolescence: A five-wave longitudinal study. *Journal of Youth and Adolescence*, *38*, 839-849.

Dekker, M. C., Ferdinand, R. F., Van Lang, N. D., Bongers, I. L., Van Der Ende, J., & Verhulst, F. C. (2007). Developmental trajectories of depressive symptoms from early childhood to late adolescence: gender differences and adult outcome. *Journal of Child Psychology and Psychiatry*, *48*, 657-666.

Edwards, M. C., & Wirth, R. J. (2009). Measurement and the study of change. *Research in Human Development*, *6*, 74-96.

Edwards, M.C., & Wirth, R.J. (2012). Valid measurement without factorial invariance: A longitudinal example. In J.R. Harring & G.R. Hancock (Eds.), Advances in longitudinal methods in the social and behavioral sciences (pp.289-311). Thousand Oaks, CA: Sage.

Eisner, M., & Ribeaud, D. (2007). Conducting a criminological survey in a culturally diverse context lessons from the Zurich project on the social development of children. *European Journal of Criminology*, *4*, 271-298.

Ferrer, E., Balluerka, N., & Widaman, K. F. (2008). Factorial invariance and the specification of second-order latent growth models. *Methodology*, *4*, 22-36.

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: a research note. *Journal of child psychology and psychiatry*, *38*, 581-586.

Gorsuch, R. L. (1983). Factor analysis. 2nd. *Hillsdale, NJ: LEA*.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*, 1-55.

Johnson, E. C., Meade, A. W., & DuVernet, A. M. (2009). The role of referent indicators in tests of measurement invariance. *Structural Equation Modeling*, *16*(4), 642-657.

Leopold, D. R., Christopher, M. E., Burns, G. L., Becker, S. P., Olson, R. K., & Willcutt, E. G. (2016). Attention-deficit/hyperactivity disorder and sluggish cognitive tempo throughout childhood: temporal invariance and stability from preschool through ninth grade. *Journal of Child Psychology and Psychiatry*, *57*, 1066–1074.

Lösel, F., & Stemmler, M. (2012). Preventing child behavior problems in the Erlangen-Nuremberg Development and Prevention Study: results from preschool to secondary school age. *International Journal of Conflict and Violence*, *6*, 214-224.

Luengo Kanacri, B. P., Pastorelli, C., Eisenberg, N., Zuffianò, A., & Caprara, G. V. (2013). The development of prosociality from adolescence to early adulthood: The role of effortful control. *Journal of Personality*, *81*, 302-312.

Malti, T., Ribeaud, D., & Eisner, M. P. (2011). The effectiveness of two universal preventive interventions in reducing children's externalizing behavior: a cluster randomized controlled trial. *Journal of Clinical Child & Adolescent Psychology*, *40*, 677-692.

Marmorstein, N. R. (2009). Longitudinal associations between alcohol problems and depressive symptoms: early adolescence through early adulthood. *Alcoholism: Clinical and Experimental Research*, *33*, 49-59.

Martino, S. C., Ellickson, P. L., Klein, D. J., McCaffrey, D., & Edelen, M. O. (2008). Multiple trajectories of physical aggression among adolescent boys and girls. *Aggressive Behavior*, *34*, 61-75.

Mathyssek, C. M., Olino, T. M., Hartman, C. A., Ormel, J., Verhulst, F. C., & Van Oort, F. V. (2013). Does the Revised Child Anxiety and Depression Scale (RCADS) measure anxiety symptoms consistently across adolescence? The TRAILS study. *International Journal of Methods in Psychiatric Research*, *22*, 27-35.

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology*, *93*, 568-592.

Millsap, R. E., & Cham, H. (2012). Investigating factorial invariance in longitudinal data In: Laursen B., Little T.D., Card N.A. (eds). *Handbook of development research methods.* The Guildford Press; New York: 2012. pp.265-278.

Motl, R. W., Dishman, R. K., Birnbaum, A. S., & Lytle, L. A. (2005). Longitudinal invariance of the Center for Epidemiologic Studies-Depression Scale among girls and boys in middle school. *Educational and Psychological Measurement*, *65*, 90-108.

Murray, A. L., Eisner, M., & Ribeaud, D. (2016). The development of the general factor of psychopathology 'p-factor' through childhood and adolescence. *Journal of Abnormal Child Psychology,* 1573–1586.

Murray, A. L., Eisner, M., & Ribeaud, D. (2016). Development and Validation of a Brief Measure of Violent Thoughts The Violent Ideations Scale (VIS). *Assessment*. Online First.

Murray, A. L., Eisner, M., & Ribeaud, D. (2017). Can the Social Behavior Questionnaire help meet the need for dimensional, transdiagnostic measures of childhood and adolescent psychopathology? *European Journal of Psychological Assessment.* In Press.

Murray, A. L., Obsuth, I., Zirk-Sadowski, J., Ribeaud, D., & Eisner, M. (2016). Developmental relations between ADHD symptoms and reactive versus proactive aggression across childhood and adolescence. *Journal of Attention Disorders*. Online First.

Murray, A. L., Obsuth, I., Eisner, M., & Ribeaud, D. (2017). Identifying early markers of later onset ADHD symptoms. *Journal of Attention Disorders.* In Press.

Nantel-Vivier, A., Kokko, K., Caprara, G. V., Pastorelli, C., Gerbino, M. G., Paciello, M., ... & Tremblay, R. E. (2009). Prosocial development from childhood to adolescence: a multi-informant perspective with Canadian and Italian longitudinal studies. *Journal of Child Psychology and Psychiatry*, *50*, 590-598.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, 111-163.

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, *17*, 354-373.

Rouquette, A., Côté, S. M., Pryor, L. E., Carbonneau, R., Vitaro, F., & Tremblay, R. E. (2014). Cohort profile: The Quebec longitudinal study of kindergarten children (QLSKC). *International Journal of Epidemiology*, *43*, 23-33.Rutter, M. (1967). A children's behaviour questionnaire for completion by teachers: preliminary findings. *Journal of child Psychology and Psychiatry*,*8*, 1-11.

Schermelleh-Engel K., Moosbrugger H. & Muller H. (2003). Evaluating the fit of structural equation models: tof significance and descriptive goodness-of-fit measures. *Methods of Psychological Research,* 8, 23-74.

Spaeth, M., Weichold, K., & Silbereisen, R. K. (2015). The development of leisure boredom in early adolescence: Predictors and longitudinal associations with delinquency and depression. *Developmental Psychology*, *51*, 1380-1394.

Sprott, J.B., Jenkins, J.M., Doob, A.N. (2000). National Longitudinal Survey of Children and Youth. Quebec, Canada: Report prepared for the Applied Research Branch, Strategic Policy, Human Resources Development Canada.

Sterba, S. K., Copeland, W., Egger, H. L., Jane Costello, E., Erkanli, A., & Angold, A. (2010). Longitudinal dimensionality of adolescent psychopathology: testing the differentiation hypothesis. *Journal of Child Psychology and Psychiatry*, *51*, 871-884.

Tremblay, R. E., Loeber, R., Gagnon, C., Charlebois, P., Larivee, S., & LeBlanc, M. (1991). Disruptive boys with stable and unstable high fighting behavior patterns during junior elementary school. *Journal of Abnormal Child Psychology*, *19*, 285-300.

Tremblay, R. E., Vitaro, F., Gagnon, C., Piché, C., & Royer, N. (1992). A prosocial scale for the Preschool Behaviour Questionnaire: Concurrent and predictive correlates. *International Journal of Behavioral Development*, *15*, 227-245

van Lier, P. A., Vitaro, F., Barker, E. D., Brendgen, M., Tremblay, R. E., & Boivin, M. (2012). Peer victimization, poor academic achievement, and the link between childhood externalizing and internalizing problems. *Child Development*, *83*, 1775-1788.

Van Oort, F. V. A., Greaves-Lord, K., Verhulst, F. C., Ormel, J., & Huizink, A. C. (2009). The developmental course of anxiety symptoms during adolescence: the TRAILS study. *Journal of Child Psychology and Psychiatry*, *50*, 1209-1217.

Verhoeven, M., Sawyer, M. G., & Spence, S. H. (2013). The factorial invariance of the CES-D during adolescence: are symptom profiles for depression stable across gender and time?. *Journal of Adolescence*, *36*, 181-190.

Rutter, M. (1967). A children's behaviour questionnaire for completion by teachers: preliminary findings. *Journal of child Psychology and Psychiatry*,*8*, 1-11.

Sass, D. A., Schmitt, T. A., & Marsh, H. W. (2014). Evaluating model fit with ordered categorical data within a measurement invariance framework: A comparison of estimators. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*, 167-180.

Weir, K., & Duveen, G. (1981). Further development and validation of the prosocial behaviour questionnaire for use by teachers. Journal of Child Psychology and Psychiatry, 22, 357-374.

Widaman, K. F., Ferrer, E., & Conger, R. D. (2010). Factorial invariance within longitudinal structural equation models: Measuring the same construct across time. *Child Development Perspectives*, *4*, 10-18.

Widaman, K. F., Little, T. D., Geary, D. C., & Cormier, P. (1992). Individual differences in the development of skill in mental addition: Internal and external validation of chronometric models. *Learning and Individual Differences*, *4*, 167-213.

Yoon, M., & Millsap, R. E. (2007). Detecting violations of factorial invariance using data-based specification searches: A Monte Carlo study. *Structural Equation Modeling*, *14*(3), 435-463.